

GISSD: Group I Intron Sequence and Structure Database

Yu Zhou^{1,2}, Chen Lu¹, Qi-Jia Wu¹, Yu Wang¹, Zhi-Tao Sun¹, Jia-Cong Deng¹ and Yi Zhang^{1,*}

¹State Key Laboratory of Virology, College of Life Sciences, Wuhan University, Wuhan, Hubei 430072, People's Republic of China and ²LRI, UMR CNRS 8623, Université Paris-Sud 11 F91405 Orsay Cedex, France

Received August 15, 2007; Revised September 7, 2007; Accepted September 11, 2007

ABSTRACT

Group I Intron Sequence and Structure Database (GISSD) is a specialized and comprehensive database for group I introns, focusing on the integration of useful group I intron information from available databases and providing *de novo* data that is essential for understanding these introns at a systematic level. This database presents 1789 complete intron records, including the nucleotide sequence of each annotated intron plus 15 nt of the upstream and downstream exons, and the pseudoknots-containing secondary structures predicted by integrating comparative sequence analyses and minimal free energy algorithms. These introns represent all 14 subgroups, with their structure-based alignments being separately provided. Both structure predictions and alignments were done manually and iteratively adjusted, which yielded a reliable consensus structure for each subgroup. These consensus structures allowed us to judge the confidence of 20085 group I introns previously found by the INFERNAL program and to classify them into subgroups automatically. The database provides intron-associated taxonomy information from GenBank, allowing one to view the detailed distribution of all group I introns. CDSs residing in introns and 3D structure information are also integrated if available. About 17000 group I introns have been validated in this database; ~95% of them belong to the IC3 subgroup and reside in the chloroplast tRNA^{Leu} gene. The GISSD database can be accessed at <http://www.rna.whu.edu.cn/gissd/>

INTRODUCTION

Group I introns are widespread non-coding RNAs found in the nuclear, chloroplast and mitochondrial genomes of

eukaryotes, in archaeabacterial and eubacterial genomes and even in some viral genomes. These introns interrupt tRNA, rRNA and mRNA genes, and their splicing from precursor RNAs is catalyzed by the intron itself, which folds to a well-defined three-dimensional structure (1–3). Group I introns are highly diverse in their lengths and primary sequences, but possess a common secondary structure, including 10 bp regions designated P1–P10, with P1 and P10 constituting the substrate domain and P3–P9 forming the compact catalytic core structure (3,4). Appendices to the conserved core elements, non-conserved peripheral elements are present for all group I introns (1). The differences in these peripheral structures and core sequences divide group I introns into 14 subgroups (4,5). It was shown that the peripheral structures co-vary with core sequences (5). Both crystal structure and biochemical studies have pointed out that intron-specific peripheral elements play a large role in organizing the tertiary structure and in encoding the distinct folding feature of each ribozyme (3,6–8). It remains to be elucidated how the highly variable peripheral elements contribute to group I intron folding and interaction with the cellular factors *in vivo*.

Group I introns are considered to be a class of selfish genetic elements because they do not seem to benefit the host organisms and are efficiently spread at the DNA level into intronless cognate sites by a process termed homing. Intron homing is catalyzed by the homing endonuclease encoded by homing endonuclease gene (HEG) residing in the peripheral elements or junctions linking adjacent core elements (9). The distribution of group I introns is highly sporadic in bacteria and lower eukaryotes, and no group I intron has been reported in humans and higher animals, supporting the view that group I introns lack an important biological function. However, in the conserved *trnT–trnF* region of the chloroplast genomes ranging from mosses to seed plants, a group I intron splitting the tRNA^{Leu} gene is a stable component of the *trnT–trnL–trnF* cistron and has been widely used for reconstructing phylogenies between

*To whom correspondence should be addressed. Tel: +86 27 68756207; Fax: +86 27 68754945; Email: yizhang@whu.edu.cn

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

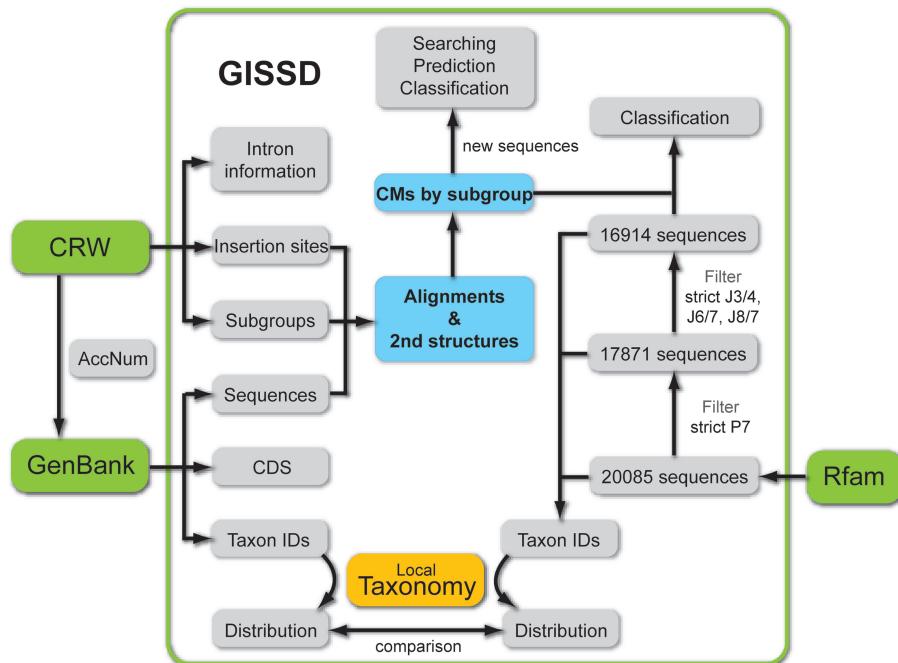


Figure 1. GISSD pipeline. Green blocks indicate foreign databases, blue blocks highlight the core data of GISSD and the orange block indicates the local taxonomy data in GISSD. CM: covariance model, which is computed from intron alignments by the INFERNAL software package.

closely related species and for identifying plant species (10–13). The tRNA^{Leu} intron found in cyanobacteria is highly homologous to those across the spectrum of chloroplasts, suggesting that it is an ancient intron present in their common ancestor (14).

As of March 2005, a total of 2800 group I introns were documented in eukaryotic genomes on the Comparative RNA Website (CRW) database (<http://www.rna.icmb.utexas.edu>) (15). This number had not increased in the newest version updated in January 2007 (<http://www.rna.cccb.utexas.edu>). Interestingly, the Rfam database reported a much larger number of group I introns (<http://www.sanger.ac.uk/Software/Rfam>) (16); as of February 2007, a total number of 20 085 group I introns are annotated in this database. Rfam predicts non-coding RNAs, including group I introns, from the public database using the INFERNAL software package (17), featuring multiple sequence alignments and profile stochastic context-free grammars, whereas CRW is an online database of comparative sequence and structure information for ribosomal, intron and other RNAs (15). It seems that CRW has its strength in annotating group I introns in rRNA genes but not in tRNA genes, wherein Rfam predicts a group I intron based on a single consensus structure derived from an alignment of 30 seed group I introns. The identification of group I introns in the CRW database is expected to be highly confident because most of them have been classified into one of 12 subgroups, while no evaluation of the confidence of group I introns in the Rfam database has ever been attempted.

GISSD is a specialized and comprehensive database for group I introns. This database aims to provide a consensus

structure for each subgroup of group I introns based on high quality alignments, to judge the confidence of the group I introns annotated by Rfam, to classify Rfam group I introns into subgroups based on those consensus structures and to provide intron number-containing taxonomy trees based on the taxonomy information of the host organisms of all group I introns. The strategy of this work is listed in Figure 1. First, starting from the GenBank accession number provided for each group I intron, the reliable secondary structures and alignments of 1789 group I introns representing 14 subgroups were obtained using a comparative sequence analysis approach (Figure 2 and Supplementary Table S1). Second, those alignments revealed distinct structure characters for P7, J6/7, J8/7 and J3/4 (Supplementary Tables S2 and S3), which were then used to judge the confidence of group I introns in Rfam. The analysis showed that 17871 of Rfam introns have a common P7 structure and 16914 introns further satisfied structural constraints for J6/7, J8/7 and J3/4. Third, we deduced a consensus structure for each subgroup according to the alignments of 1789 introns, which was then processed by INFERNAL software package to classify the 16914 confident group I introns into subgroups, with 16146 introns being classified into the IC3 subgroup. Fourth, we extracted taxonomic information associated with the organism harboring each group I intron from GenBank, and present a taxonomy-based phylogeny tree of group I introns. It is shown that 16299 of group I introns reside in the genomes of Viridiplantae. Given that the intron splitting the chloroplast tRNA^{Leu} genes belongs to IC3 subgroup, these results strongly suggest that a major reservoir of group I introns is the chloroplast tRNA^{Leu} genes.

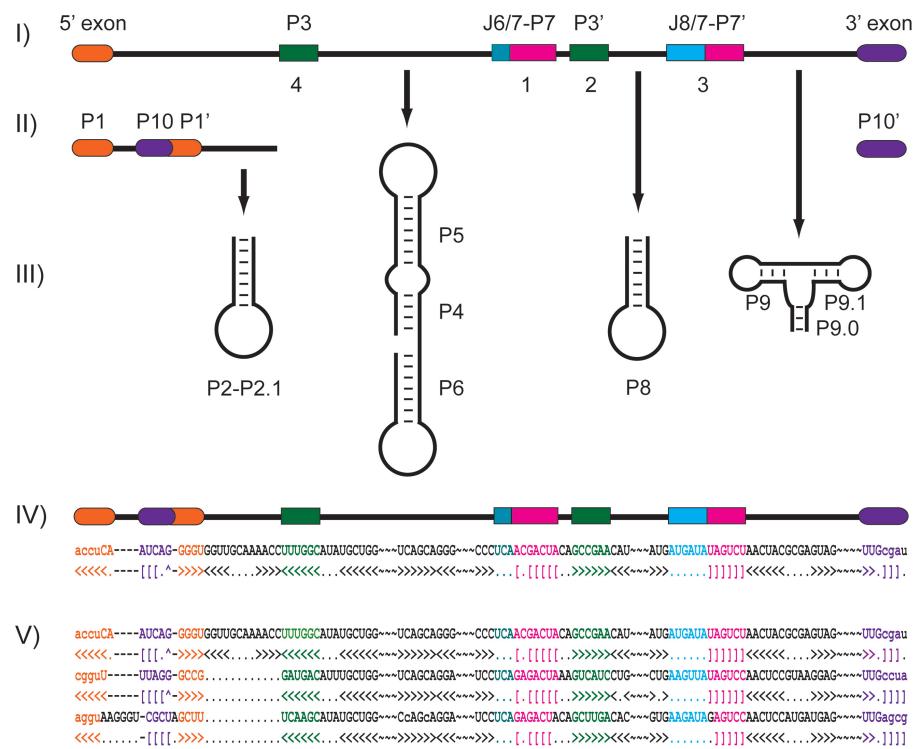


Figure 2. The schematic representation of secondary structure prediction and alignment. (I) Locate the core components in the intron by using conservative sequence patterns. The order is as: (1) find J6/7-P7, which was very conservative; (2) find P3', which is usually 2 or 3 nt after P7, if no insertion sequences; (3) find J8/7-P7', according to the base pairing of P7 and P7' and the conservation of J8/7 and (4) find P3, paired with P3'. (II) Partition the sequence into four parts. In the first part, 5' exon and 3' exon sequences were used to find P1' and P10, and the rest of the sequence before P3 was used to identify P2 and P2.1. (III) The four parts were folded by *RNAstructure* 4.11 (18) separately. Besides the minimum energy structure, other suboptimal structures were also checked. By comparing the folded structure to known structures, the structure having similar pattern was chosen manually. (IV) The whole structure of an intron was completed by integrating the structures of the subsequences. (V) When structure prediction was finished for certain numbers of introns in a subgroup, the alignment process was started. The core components and peripheral elements were sequentially aligned manually based on their structures. A point to emphasize is that the aligning process and the structure prediction procedure were iteratively done. Once one part of an intron ran into difficulty in the alignment with other sequences, the corresponding structure was reselected from the candidate structures of *RNAstructure*. Sometimes the core components needed to be reconsidered, and the whole process of structure prediction was redone for that particular intron.

OVERVIEW OF THE DATA

Sequence full information

Among the 1921 group I introns that have been assigned to subgroups on the CRW database (as of 16 March 2005), only 1829 intron sequences were retrieved from GenBank according to the accession number provided by CRW and the intron annotation in the GenBank file (Supplementary Table S1). Two alternative approaches were used to identify introns when the annotation was not available. The full sequence in GenBank file was aligned with the mature rRNA from the closest species. Alternatively, the 5' exon and 3' exon of the intron were identified by matching the full sequence with the characteristic exon sequence from the same insertion position. Introns of the IE major subgroup have been classified to the IE1, IE2 and IE3 minor subgroups, according to recently published work (5); adding this new classification information made the final 14 subgroups (Supplementary Table S1). The GenBank accession number, subgroup, insertion position for rRNA introns, host gene name and type, host organism and cellular location of each group I intron were from CRW.

For the convenience of referral, each intron was assigned a unique name. The nomenclature of the rRNA introns is according to the standard proposed by Johansen and Haugen (19) with a few exceptions to avoid duplicate names. We hope that this can provide an opportunity for the community to adopt a unified nomenclature for group I introns.

Mistakes in the intron information obtained from GenBank and CRW were found and corrected; the correction is marked under the correction part of the corresponding intron. Exons of a small number of introns used for structure prediction were not available in the GenBank records, and thus homologous sequences from the species closest in taxonomy were borrowed; this information is placed under the description part of the corresponding intron.

The taxonomic information was downloaded from the NCBI Taxonomy FTP site (<ftp://ftp.ncbi.nih.gov/pub/taxonomy/>). PDB IDs of the 3D structure related to group I introns were from Protein Data Bank (<http://www.pdb.org>). The related information of the CDSs in group I introns, including start position in intron, length, name and protein ID (if available), were recorded in GISSD.

Table 1. Intron distribution on the host phylogenetic tree

Phylogeny of Host Genes	Intron Numbers				Phylogeny of Host Genes	Intron Numbers			
	Rfam Original	Strict P7	Strict P7, J34, J67 and J87	CRW		Rfam Original	Strict P7	Strict P7, J34, J67 and J87	CRW
0-> root	20085	17871	16914	1789					
1-> Viruses	31	31	20	13					
2-> dsDNA viruses, no RNA stage	31	31	20	13					
3-> Phycodnaviridae	7	7	7	3					
3-> Caudovirales	24	24	13	10					
1-> cellular organisms	20054	17840	16894	1776					
2-> Bacteria	394	392	68	90					
3-> Cyanobacteria	383	381	57	73					
3-> Proteobacteria	5	5	5	6					
3-> Firmicutes	6	6	6	2					
3-> Chlamydiae/Verrucomicrobia group 0	0	0	0	1					
3-> Thermotogae	0	0	0	8					
2-> Eukaryota	19660	17448	16826	1686					
3-> Rhodophyta	5	2	1	144					
3-> Heterolobosea	11	9	9	22					
3-> Granuloreticulosea	0	0	0	1					
3-> Viridiplantae	18998	16863	16299	717	3-> Viridiplantae	18998	16863	16299	717
3-> Fungi/Metazoa group	579	517	465	755	4-> Chlorophyta	252	241	211	313
3-> Alveolata	16	13	12	13	5-> Prasinophyceae	4	2	1	15
3-> stramenopiles	6	5	5	5	5-> Chlorophyceae	124	118	99	178
3-> Acanthamoebidae	4	4	4	14	5-> Ulvophyceae	35	32	27	16
3-> Euglenozoa	1	1	1	1	5-> Pedinophyceae	3	3	2	3
3-> Plasmidiophorida	14	14	14	2	5-> Trebouxiphycaceae	85	85	81	101
3-> Glauco cystophyceae	4	4	4	1	5-> unclassified Chlorophyta	1	1	1	0
3-> Lobosea	1	1	1	0	4-> Streptophyta	18746	16622	16088	404
3-> Cercozoa	2	2	1	0	5-> Mesostigmaphyceae	1	1	0	4
3-> Mycetozoia	16	10	7	11	5-> Zygnemophyceae	18	17	16	105
3-> Arcellinida	3	3	3	0	5-> Chlorokybophyceae	1	1	1	0
					5-> Streptophytina	18726	16603	16071	291

The phylogenies of host organisms are listed on the left and the corresponding numbers of introns annotated in Rfam and CRW are listed in the next four columns. Introns from Rfam and those being filtered using different restrictions are also indicated. The shaded Viridiplantae region representing the major intron difference between Rfam and CRW are extended for two more nodes on the right.

Besides the CDSs annotated in GenBank records, all other introns longer than 600 nt were subjected to a BLASTX search (<http://www.ncbi.nlm.nih.gov/BLAST/>) with an *E*-value cutoff of 10^{-4} to identify the presence of CDSs. Intron-encoded homing endonuclease and maturase were specified.

Secondary structure and alignment

The methods used here are similar to those previously reported (5); the structure prediction procedure and the aligning process were iteratively done to guarantee the accuracy (Figure 2). Introns in a major subgroup or minor subgroup were assigned to a curator for prediction, and the initial structures and alignments were checked, validated or revised by a senior curator to ensure the quality and accuracy of the final ones presented to the society. The figure of the secondary structures including two pseudoknots (P3–P7, P1–P10) was generated by RnaViz 2.0 (20), followed by manual adjustments to obey the three-domain presentation of a group I intron structure. A new structural alignment format for group I introns was defined to keep maximum structure information in the alignment (Supplementary Figure S1).

Processing of group I introns from Rfam (gIRfam)

Rfam predicts group I introns based on a single consensus structure derived from an alignment of 30 seed group I introns (16). Twenty-six of these seed introns are included in the group I intron dataset analyzed above, with 7, 6, 5, 3 and 2 of the seeds belonging to IC3, IC2, IA1, IC1 and IB3, respectively. There is only one sequence each for subgroups IA2, IA3 and IB4, and no sequence is present

for IB1, IB2, ID and IE in the seed alignment. The consensus structure has considered the eight conserved base-paired elements, including P1–P6 and P8–P9. P7 forms a pseudoknot with P3, and therefore has been excluded from the INFERNAL search. The secondary structure-based alignments of 1789 group I introns belonging to 14 different subgroups revealed distinct structure constraints for P7 (Supplementary Table S2), as well as length constraints for the conserved joint regions J6/7, J8/7 and J3/4 (Supplementary Table S3). Searching all 20085 introns present in Rfam showed that 17871 of Rfam introns satisfy P7 constraints and ~11% Rfam introns lack a typical P7 structure. Among the P7-containing introns, 16914 contain restricted J6/7, J8/7 and J3/4 (Table 1). These results suggest that INFERNAL provides a relatively reliable method in the *de novo* search of group I introns from sequence databases (89% containing restricted P7 and 94.6% of which also containing restricted joint regions). Therefore, the number of group I introns in the current sequence database should be much larger than those presented in CRW.

To categorize these 16914 group I introns, we deduced a consensus structure for each of the 14 subgroups according to the alignments of the 1789 group I introns. The consensus structures were then processed by the INFERNAL software package (17). The specificity and sensitivity of these consensus structures in identifying group I introns in each subgroup was tested (Supplementary Tables S4 and S5). We chose a cutoff with a strict specificity to select group I introns for each subgroup, showing that 16146 introns belong to the IC3 subgroup, representing 95.5% of all 16914 group I introns. The next large subgroup revealed by this classification is IC1,

containing 428 introns (2.53%). IC1 subgroup mainly contains group I introns that interrupt rRNA genes and represents the largest subgroup in the CRW database (49.2%), whereas IC3 introns only represent 18.6% of group I introns in this rRNA-focused database. A total of 16 729 group I introns have been categorized into 10 subgroups, representing 98.9% of all input group I introns (Supplementary Table S6). Only one ID intron was identified and no IE intron was found, which is consistent with the fact that no ID and IE introns are present in the seed alignment.

Group I intron distribution: the lineage of the host organisms containing group I introns

A taxonomy ID (taxonid) is assigned to each taxon (species, genus, family, etc.) in the NCBI Taxonomy Database (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=Taxonomy>). This database also contains the information of the parent-child relationship of those taxonids. Based on this database and the taxonids of the host organisms containing group I introns, the number of group I introns belonging to each taxonid was computed and an intron number-containing taxonomy tree was constructed. The number of group I introns belonging to each taxonid was pre-computed and stored in our database to guarantee the speed of the tree construction during each viewing.

The intron-containing taxonomy tree reveals that 16 299 (96.4%) of the 16 914 confident group I introns reside in the genomes of Viridiplantae, whereas only 717 (37.3%) of the 1921 introns in the CRW database reside in Viridiplantae (Table 1). Because most of the Viridiplantae introns reside in the tRNA^{Leu} gene of the chloroplast genome, this result suggests that the CRW database is very inefficient in annotating tRNA group I introns. However, a larger number of introns from the Fungi/Metazoa group was annotated by CRW than by Rfam (755 versus 465), indicating that CRW is more powerful in annotating fungal rRNA introns.

Summary of the major finding

It is very interesting to find that group I introns are prevalent in Viridiplantae and the IC3 subgroup. Considering that the chloroplast tRNA^{Leu} belongs to IC3, data provided by GISSD strongly suggest that the major reservoir of group I introns in nature lies in the chloroplast tRNA^{Leu} genes.

IMPLEMENTATION AND WEB INTERFACE

GISSD is a relational database implemented with MySQL 5.0.15 on a Redhat 9.0 Linux system running on a HP ML150 server. A user-friendly Web interface was developed for easy viewing and retrieval by CGI scripts. Figure 3 shows the navigation of the Web pages:

- Home page gives the overview of GISSD and the background of group I introns.
- Search page allows the user to query all introns in a chosen subgroup directly, or to perform a limited

search using the provided form. The search result page lists hits of the query by line displaying important information, including intron name, subgroup, organism name, insertion position and links to intron sequence page, secondary structure and the GenBank record. The intron sequence page displays the full intron information in a well-defined format, including unique intron name, GenBank accession number, host organism, host lineage, host gene name, host gene type, the cellular location of the host gene, intron classification, strand being transcribed, insertion position of the rRNA introns, the intron length, the intron sequence in Fasta format and exon sequences. The correction information, description of more details, CDS and 3D structure information are given if available.

- Sequence page provides gzipped and zipped Fasta files containing all entries in each subgroup for download. Upstream and downstream 15 nt exon sequences are also included, which are in lowercase, and intron sequences are in uppercase.
- Structure page provides tar-gzipped and zipped file containing all PDF structures files in each subgroup for download.
- Alignment page shows 11 novel alignments of subgroups IA1, IA2, IA3, IB1, IB2, IB3, IB4, IC1, IC2, IC3, ID and three published alignments of IE1, IE2, IE3 (5). An alignment of 61 unclassified IE introns is also provided. All of these alignments are available for free download.
- Distribution page allows a user to display the desired level of taxonomy nodes. In each taxonomy node, the level, taxonomy name and rank and the number of introns with sequence and structure are shown. Clicking on the last taxonomy node displayed in the top lineage line will display all the introns in this node in a search result format.
- gIRfam page presents information for all group I introns in Rfam, evaluation and classification results of these introns and related intron-containing taxonomy tree.
- Submission page welcomes users to submit new and updated information of group I introns, which will be validated and analyzed by our curators and for which the final results will be loaded to the database and sent back to the submitter.

FUTURE PLANS

GISSD is a specialized and comprehensive database for group I introns, focusing on integrating useful group I intron information from all available databases. GISSD also aims to provide *de novo* data essential for understanding group I introns at a systematic level. Upcoming tasks include developing an automatic program to predict the secondary structure, annotating new group I introns and categorizing group I introns. New information will be constantly added to the database to provide the RNA community the most updated scenario of group I intron study.

A

Please choose the subgroup(s) in the left hierarchy tree!

Intron name: Host gene name: Organism: Phylogeny: Accession number: Insertion position: Length: Cell location: Host gene type: CDS: 3D structure: Sort by: Records/Page: View introns Reset

B

The result of your query: total record: 837 , total page: 42

Num	Intron name	Subgroup	Organism	IP	Length	2nd struct.	AccNum
1	Ac.S943	IC1	Apioagladiostoma aceriferum	943	438	✓	AF281664
2	Aad.L2449	IC1	Arcula adenitrovirus	2449	411	✓	Z50840
3	Aau.S943	IC1	Artemyces austropiperatus	943	403	✓	AY298775
4	Abo.S1506	IC1	Arthrobotrys botryospora	1506	428	✓	US1955

C

No.	Subgroup	Number of introns	Download
1	IA1	80	[IA1_seq.fasta.gz] [IA1_seq.fasta.zip]
2	IA2	15	[IA2_seq.fasta.gz] [IA2_seq.fasta.zip]
3	IA3	56	[IA3_seq.fasta.gz] [IA3_seq.fasta.zip]
4	IB1	42	[IB1_seq.fasta.gz] [IB1_seq.fasta.zip]
5	IB2	18	[IB2_seq.fasta.gz] [IB2_seq.fasta.zip]
6	IB3	7	[IB3_seq.fasta.gz] [IB3_seq.fasta.zip]
7	IB4	89	[IB4_seq.fasta.gz] [IB4_seq.fasta.zip]
8	IC1	837	[IC1_seq.fasta.gz] [IC1_seq.fasta.zip]
9	IC2	32	[IC2_seq.fasta.gz] [IC2_seq.fasta.zip]
10	IC3	328	[IC3_seq.fasta.gz] [IC3_seq.fasta.zip]
11	ID	17	[ID_seq.fasta.gz] [ID_seq.fasta.zip]
12	IE1	38	[IE1_seq.fasta.gz] [IE1_seq.fasta.zip]
13	IE2	58	[IE2_seq.fasta.gz] [IE2_seq.fasta.zip]
14	IE3	111	[IE3_seq.fasta.gz] [IE3_seq.fasta.zip]
15	IE	61	[IE_seq.fasta.gz] [IE_seq.fasta.zip]

D

Please select the maximum level to view per page: 3 View

Current lineage: Hit the last node in red to see the introns belonging to it, Hit other nodes, go back to upper level root

- 0-> root (no rank) (1700)
 - 1-> Viruses (no rank) (13)
 - 2-> dsDNA viruses, no RNA stage (no rank) (13)
 - 3-> Phycodnaviridae (family) (3)
 - 3-> Cauloviridae (order) (10)
 - 1-> cellular organisms (no rank) (1776)
 - 2-> Bacteria (superkingdom) (90)
 - 3-> Cyanobacteria (phylum) (73)
 - 3-> Proteobacteria (phylum) (6)
 - 3-> Actinobacteria (phylum) (2)
 - 3-> Chlamydias/Verrucomicrobia group (superphylum) (1)
 - 3-> Thermotoga (phylum) (8)
 - 2-> Eukaryota (superkingdom) (1886)
 - 3-> Rhizophyta (no rank) (144)
 - 3-> Heterolobosea (class) (22)
 - 3-> Gramicolae/reticulosea (class) (1)
 - 3-> Viridiplantae (kingdom) (717)
 - 3-> Fungi-Metazoa group (no rank) (755)

E

Classification

Classification of confident Rfam group I introns (16914) into subgroups using the MER threshold (cutoff) obtained in 5-fold cross validation. A very small number of introns were not well resolved. A total of 35 group I introns were classified into two subgroups, 8 introns into three subgroups, and 2 into four subgroups. Introns in IB subgroups are particularly poorly resolved probably because their high structure similarity.

A tab-delimited file containing the subgroup information: Download

Subgroup seeds	Cut-off score	Positive hits	Hit one of the other subgroups	Hit two of the other subgroups	Hit three of the other subgroups
IA1	69.76	64			
IA2	86.84	9	1		
IA3	49.09	3	3		
IB1	98.68	15	1	8	2
IB2	89.70	32	21	7	2
IB3	68.00	5		1	2
IB4	76.15	45	20	8	2
IC1	26.40	428	11		
IC2	108.29	39	11		
IC3	41.53	16146	2		
ID	39.41		1		
IE1	195.05	0			
IE2	218.30	0			
IE3	151.58	0			

Figure 3. Screenshots of GISSD. (A) Intron search page, (B) search result page, (C) sequence, structure and alignment page, (D) intron distribution page and (E) gIRfam page.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Ross Fu (Caltech), Jessica Shin (Stanford University) and Prof. Alain Denise (University of Paris-Sud) for their careful reading of the manuscript. We also thank Feiyang Cai and Prof. Zhengyou Liu (Department of Physics, Wuhan University) for timely providing a computer cluster to process Rfam group I introns. This work is supported by the National Natural Science Foundation of China (30330170 and 90608025) and by the National Basic Research Program of China (2005CB724604), granted to Y.Z. Funding to pay the

Open Access publication charges for this article was provided by grant 90608025 from The National Nature Science Foundation of China.

Conflict of interest statement. None declared.

REFERENCES

- Burke,J.M., Belfort,M., Cech,T.R., Davies,R.W., Schwelen,R.J., Shub,D.A., Szostak,J.W. and Tabak,H.F. (1987) Structural conventions for group I introns. *Nucleic Acids Res.*, **15**, 7217–7221.
- Cech,T.R. (1990) Self-splicing of group I introns. *Annu. Rev. Biochem.*, **59**, 543–568.
- Vicens,Q. and Cech,T.R. (2006) Atomic level architecture of group I introns revealed. *Trends Biochem. Sci.*, **31**, 41–51.
- Michel,F. and Westhof,E. (1990) Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.*, **216**, 585–610.

5. Li,Z.J. and Zhang,Y. (2005) Predicting the secondary structures and tertiary interactions of 211 group I introns in IE subgroup. *Nucleic Acids Res.*, **33**, 2118–2128.
6. Rangan,P., Masquida,B., Westhof,E. and Woodson,S.A. (2003) Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme. *Proc. Natl Acad. Sci. USA*, **100**, 1574–1579.
7. Xiao,M., Leibowitz,M.J. and Zhang,Y. (2003) Concerted folding of a Candida ribozyme into the catalytically active structure posterior to a rapid RNA compaction. *Nucleic Acids Res.*, **31**, 3901–3908.
8. Woodson,S.A. (2005) Metal ions and RNA folding: a highly charged topic with a dynamic future. *Curr. Opin. Chem. Biol.*, **9**, 104–109.
9. Haugen,P., Simon,D.M. and Bhattacharya,D. (2005) The natural history of group I introns. *Trends Genet.*, **21**, 111–119.
10. Borsch,T., Hilu,K.W., Quandt,D., Wilde,V., Neinhuis,C. and Barthlott,W. (2003) Noncoding plastid trnT-trnF sequences reveal a well resolved phylogeny of basal angiosperms. *J. Evol. Biol.*, **16**, 558–576.
11. Quandt,D. and Stech,M. (2005) Molecular evolution of the trnL(UAA) intron in bryophytes. *Mol. Phylogenetic Evol.*, **36**, 429–443.
12. Won,H. and Renner,S.S. (2005) The chloroplast trnT-trnF region in the seed plant lineage Gnetales. *J. Mol. Evol.*, **61**, 425–436.
13. Taberlet,P., Coissac,E., Pompanon,F., Gielly,L., Miquel,C., Valentini,A., Vermaat,T., Corthier,G., Brochmann,C. and Willerslev,E. (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.*, **35**, e14.
14. Kuhsel,M.G., Strickland,R. and Palmer,J.D. (1990) An ancient group I intron shared by eubacteria and chloroplasts. *Science*, **250**, 1570–1573.
15. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D’Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Müller,K.M. et al. (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 15.
16. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
17. Nawrocki,E.P. and Eddy,S.R. (2007) Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.*, **3**, e56.
18. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
19. Johansen,S. and Haugen,P. (2001) A new nomenclature of group I introns in ribosomal DNA. *RNA*, **7**, 935–936.
20. De Rijk,P., Wuyts,J. and De Wachter,R. (2003) RnaViz2: an improved representation of RNA secondary structure. *Bioinformatics*, **19**, 299–300.